Multi-task Models for Predicting the Progression of Alzheimer's from MRI

Rebecca Zhang

Advisor: Professor Han Liu

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science in Engineering

Department of Operations Research and Financial Engineering

Princeton University

June 2015

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

_____

Rebecca Zhang

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

_____

Rebecca Zhang

Abstract

Alzheimer's disease (AD) is a severe neurodegenerative disorder affecting over 5 million in the U.S. today. By 2050, the total is estimated to reach 13.8 million. Researchers are scrambling to find methods for accurate and early prediction of AD progression, since its irreversible symptoms are particularly destructive and no treatments have been found. Previous studies have shown that changes in certain regions of the brain, as measured using magnetic resonance imaging (MRI), have been associated with the onset of AD. Clinical diagnosis of AD often relies on cognitive measures such as the Mini Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog). In this thesis, we address the problem of predicting MMSE and ADAS-Cog scores at future time points from high-dimensional MRI data collected at a baseline initial visit. We structure the problem as a multi-task regression problem, where each task is the prediction at a certain time point. We propose three novel calibrated multi-task formulations and conduct experiments to evaluate their performance, using two separate data sets from the Alzheimer's Disease Neuroimaging Initiative (ADNI), and compare the results to those of their non-calibrated counterparts, as well as those of the single-task ridge and Lasso regressions. Multi-task methods consistently outperform single-task methods, and of those, the calibrated formulations achieve near-identical performance to those that are non-calibrated. Overall, the best model for the prediction of AD progression is not always intuitive but dependent on the input data.

Acknowledgements

First and foremost, I must thank my advisor, Professor Han Liu, who not only introduced me to the rich field of statistical research of MRI data but also taught me so much about statistics and machine learning in a big data setting. I am also eternally grateful to Yuan Cao, who spent countless hours with me going over papers, data, and code. Daniel Jang and Daniel Lacker provided much assistance with their Senior Thesis Writers Group, and Haixiao Du, Yuan Liu, and Ziwei Zhu were also incredibly generous with their willingness to help throughout the research process.

I'd like to thank a number of other members of Princeton ORFE: Patrick Rebeschini for giving me life advice, Professor Amirali Ahmadi for being an inspiration when it comes to learning and research, and Professor Erhan Cinlar, who has supported me in all my endeavors during the last four years. What truly makes this department special is having the chance to befriend amazing peers. Austin, Megan, Pal, Saumya, Janie, Amanda, Kecy: I'll never forget our pset sessions and our countless inside jokes. And Ryan: I am truly surprised (but so glad!) that we aren't sick of each other yet.

I wouldn't even have the chance to thank the people above had it not been for the support of my parents, whose innumerable sacrifices got me to where I am today. Luckily, when I left home for Princeton, I found a family here as well. Boys of Patton 41: thank you for being the older brothers I've always wanted to have. Sharon, Kaitlyn, and Xinyi: you girls are going to do so much for the world, and I'm lucky to call you my friends. To my Kappa sisters, especially Stef, Ariana, Steph, Hana and Sewheat - I don't know how I would've made it through thesis-writing without you (your gifts of food and kind words sustained me during many a late night). And finally, to Sub, my best friend: thank you for all the adventures we've shared and for all the good that you've brought into my life. Never stop pwming.

For Mom, Dad, and Henry, who instilled in me the concept of lifelong learning,
way before it became the topic of this thesis

Contents

Abstract iii

Acknowledgements iv

List of Tables viii

List of Figures ix

1 Introduction **1**

Background and Motivation 1

Previous Studies 4

Multi-task Learning 6

2 Data **10**

3 Methodology and Models **14**

Ridge 15

Lasso 16

Temporal Group Lasso 18

Convex Fused Sparse Group Lasso 19

Calibrated Multi-task 20

4 Results **24**

Part I: ADNI1 Study Replication 25

Part II: Expansion to Ongoing ADNI Study Data 27

Part III: A Novel Approach 29

5 Discussion **33**

6 Conclusion **35**

A Code **37**

Bibliography **46**

List of Tables

2.1 Summary of study data 12

4.1 Performance measures from Zhou et al. (2013) 25

4.2 Performance measures for Part I results. 26

4.3 Performance measures for Part II results 27

4.4 Performance measures for Part III results 30

4.5 Comparison of calibrated versus non-calibrated TGL 31

4.6 Comparison of calibrated versus non-calibrated cFSGL 32

List of Figures

1.1 Biomarkers as indicators of dementia 3

3.1 Ridge vs lasso regression 17

3.2 A comparison of models 21

4.1 Scatter plots of actual versus predicted MMSE scores on testing data 28

4.2 Scatter plots of actual versus predicted ADAS-Cog 13 scores on testing data 29

Chapter 1

Introduction

Background and Motivation

Alzheimers disease (AD) is an irreversible neurodegenerative disease that affects an estimated 5.3 million within the United States, where it is also the 6th leading cause of death and accounts for between 60 to 70% of age-related dementia. Since 2000, while deaths from other diseases decreased significantly, deaths from AD increased by more than 70%. Because of the increasing number of individuals in the 65 or older age group in the United States, the annual number of new cases of AD is projected to double by 2050, when the total is estimated to reach 13.8 million. Another reason these figures show a rapid increase in the prevalence of the disease is the fact that no prevention methods or cures have been discovered. In addition to the psychological and emotional toll AD takes on the patients and their families, its pervasiveness in society as a whole poses an enormous financial burden. In 2015, AD and other dementias are expected to cost the country $226 billion, and that number could rise to $1.1 trillion by 2050 (Association, 2015).

As such, it is especially crucial for countering AD that evaluating an individuals risk and, if necessary, diagnosis and predicting progression of the disease all take

place early. Current diagnosis of AD relies mostly on documenting the mental decline of patients. Those who are diagnosed in this way have often already suffered severe brain damage by the time their diagnosis is final. A research goal is to discover a way to detect AD before its destructive symptoms begin, and a promising possibility is through the use of biomarkers. A biomarker, short for biological marker, is a substance, measurement or indicator of a biological state or disease. They may exist before clinical symptoms occur, and their behavior can help us to not only diagnose but also better understand AD and how it progresses in patients. Before it can be used in clinics, however, a biomarker must be validated by multiple studies of large groups of people, verifying that it can accurately and reliably indicate the presence of disease. Once a biomarker is deemed sensitive enough to detect early mild cognitive impairment or AD, it can also aid in the development and evaluation of new treatments, should they arise.

Figure 1.1 shows a graph of the behavior of five not-yet-validated but promising biomarkers as indicators of dementia. The last two curves representing memory loss and general cognitive decline respectively, as measured by cognitive assessment, reach appreciable magnitudes once full-on dementia has set in, and are the classic indicators of dementia. In fact, since a definitive diagnosis of AD requires analysis of brain tissue done during a brain biopsy or autopsy, a clinical diagnosis of probable AD often relies principally on clinical scores measured by these cognitive assessments as an important criterion. In particular, the Mini Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) were designed specifically to evaluate the cognitive status of patients, and have been shown to be a good proxy for AD progression.

In an oft-referenced study by Petrella et al. (2003), MMSE scores were shown to correlate with the underlying disease pathology of Alzheimer's as well as with the deterioration of patient's functional ability over time. They were able to determine

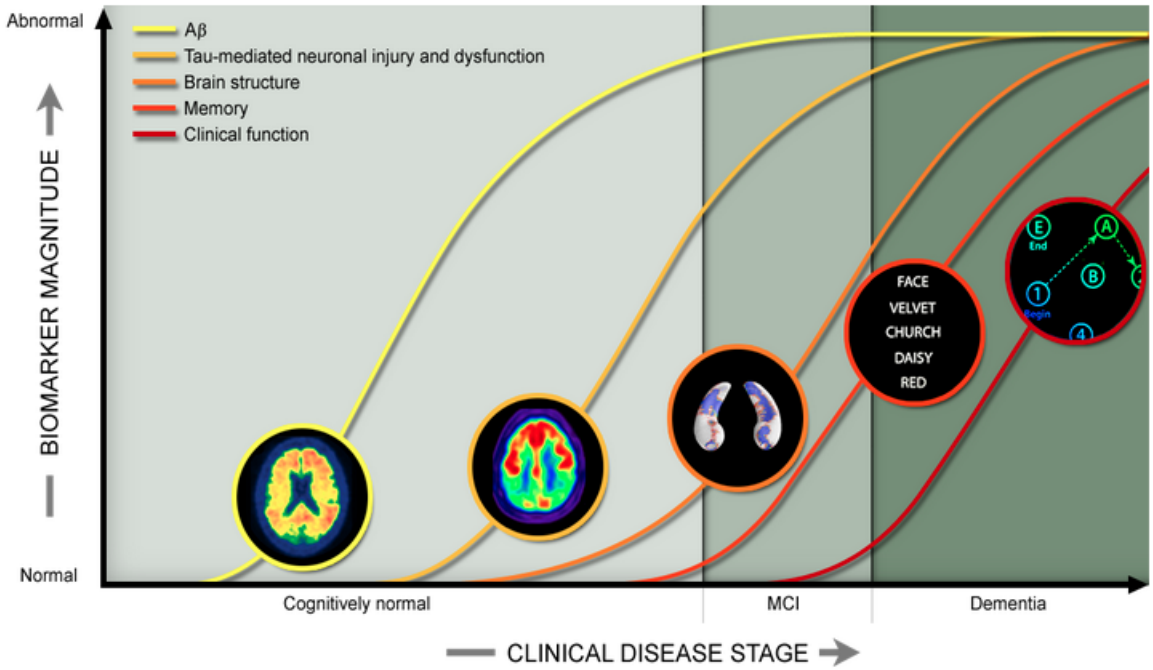

Figure 1.1: Five studied biomarkers as indicators of dementia. From left to right, these are: amyloid beta imaging detected in CSF and PET amyloid imaging; neurodegeneration detected by increase in CSF tau species and synaptic dysfunction, measured via FDG-PET; brain atrophy and neuron loss measured with MRI; memory loss measured by cognitive assessment; general cognitive decline measured by cognitive assessment.

ranges of scores that clearly associated with certain stages of the disease. Patients with MMSE score greater than 23 demonstrated minimal impairment and functioned normally and independently. When MMSE scores dropped to between 20-23, patients would exhibit forgetfulness and impairment in their daily function, indicating the onset of mild AD. Patients with moderate AD - suffering from short-term memory loss and struggling with verbal fluency - had an MMSE score of 10-19. Finally, in patients whose scores declined to below 10, the severe stage of AD was reached, indicated by behavioral changes such as delusion and aggression. Patients in this score range would eventually become completely dependent on others in most aspects of daily life.

ADAS-Cog, meanwhile, is the widely used standard for assessing cognitive function when testing the efficacy of AD drugs in their trial stages. It is a longer exam than the MMSE, requiring a higher level of training in order to administer,

but also believed to be more thorough. Therefore, in our study, we use MMSE and ADAS-Cog scores as our target values; they serve as a numerical measure of the advancement of AD, and as the response variables we wish to predict.

The first three curves in Figure 1.1 indicate changes that can be observed prior to diagnosis of dementia. All three changes are detected through neuroimaging, making it one of the most promising research areas for early detection. In particular, magnetic resonance imaging (MRI) can reveal tumors, evidence of a stroke, damage from head trauma and fluid buildup in the brain. Additionally, changes in brain atrophy and neuron loss measured by MRI lie on the border between normal cognition and the early stages of possible dementia, and this balance between reliability and timeliness makes it especially promising for predicting AD. For this reason, we choose MRI features as our input data for the prediction of MMSE and ADAS-Cog scores in our study.

Previous Studies

A number of previous studies have noted the relationship between MRI features and cognitive scores. Frisoni et al. (2002) focused specifically on the use of MRI to detect grey matter in parts of the brain. They compared 29 patients who met the conditions for probable AD with 26 normal controls, and were the first to study the correlation between MMSE and atrophy, for which they used regional grey matter density extracted through MRI as a proxy. For the AD patients, they found a positive relationship between MMSE scores and grey matter density, for both the left and right regions of the brain. Chetelat and Baron (2002) found that while most structural imaging studies have focused on the hippocampal region, in fact it was hippocampal atrophy in combination with temporal neocortex atrophy that was the most significant predictor of progression to AD. As such, they displayed a need for longitudinal

studies testing the predictive power of MRI taken at the initial visit (known as the baseline) for future possible AD conversion. Thompson et al. (2004) looked beyond the hippocampus and focused instead on the shape of ventricles. They found that baseline morphology of the brain's ventricles were closely linked with MMSE scores. While hippocampal loss rates may decline and plateau as the disease runs its course, the ventricles continue to rapidly expand, so their volume and change rates link well with MMSE decline rates. This discovery is important, as prediction of cognitive decline becomes a harder problem the further into the future one aims to predict.

It is this problem of future prediction that has been the focus of more recent studies. Duchesne et al. (2009) proposed what they called an "automated analysis technique" to find relationships between baseline MRI and the change in MMSE scores from baseline to a year after. They built a statistical model by first utilizing principle component analysis to project the high dimensional image data from a reference group into a lower dimensional space, then determining a restricted set of features highly correlated with one year change in MMSE by using bootstrapping sampling estimation, including only features that passed some predetermined threshold. From this set, they built a linear model using iterative weighted least squares regression, and found that volume and intensity of the medial temporal lobe along with age, gender, years of education and baseline MMSE scores were highly correlated with one-year changes.

MMSE and ADAS-Cog both were targets in a study by Stonnington et al. (2010). Using relevance vector regression, in which a kernel matrix is generated from image data and the probability of the clinical scores given the data is measured, they were able to predict individual's baseline performances on MMSE and ADAS-Cog (among other established cognitive assessments) from their baseline structural MRI. Their results yielded a high correlation between ground truth and predicted target value, which along with the squared error between the two are often used

as measures of model performance. They also found that MMSE and ADAS-Cog were the assessments that correlated particularly well with whole brain grey matter changes.

While the studies described above considered a large set of input features and linear models, Ito et al. (2011) proposed two power functions to model the rate of disease progression. The only variables of input were age, years of education, and baseline ADAS-Cog scores for the model utilizing numerical data, and age, gender, and APOE $\epsilon 4$ genotype for the model using categorical data. Model building and feature selection were done by adding covariates to the proposed model in a stepwise manner, and evaluating the resultant change in objective function and precision of the estimate. They found that age, APOE $\epsilon 4$ genotype, gender and baseline ADAS-Cog scores significantly affected rates of disease progression.

Rather than work with only baseline MRI, Murphy et al. (2010) focused on neuroanatomical change, since neurodegeneration can precede the onset of dementia by many years. They measured the six-month atrophy of regions in the medial temporal lobe, and used multivariate regression to determine the relationship between this change and memory reduction over two years as measured by neuropsychological exam scores. To determine which features to include, they first performed univariate analysis separately on each and included those that exceeded a set significant threshold. They found that change in three particular subregions, the right fusiform gyrus, inferior temporal cortex, and right inferior lateral ventricle, were significant predictors of future cognitive decline.

Multi-task Learning

Most existing studies focus on the prediction of AD progression status at a single particular time point (Duchesne et al., 2009; Stonnington et al., 2010; Murphy

et al., 2010). Given that the biomarkers observed through MRI are among the most promising for early AD detection, and that the process of MRI data collection is on the more labor-intensive side, we would like to find models to extract as much information from baseline MRI as possible. Namely, we would like extend the prediction problem from one time point to multiple time points, to give a possible trajectory for how AD may progress for a given patient far into the future. Such a model would be very powerful if it could both look far ahead and maintain prediction accuracy, all from just an initial screening.

In fact, under the multi-task learning model, jointly predicting cognitive scores from multiple time points is in theory expected to improve performance, especially in high dimensional cases such as this were the number of features is large and number of observations small. The idea comes from the transfer of knowledge in the lifelong learning framework, which has been shown to generalize consistently more accurately from less training data (Thrun, 1996). Consider the classic problem of object recognition from images. Suppose there are n tasks, each being to recognize a different object: a bottle, a hammer, a book, and so on, these images making up a support set. Let the n^{th} task be to recognize a shoe, on a training set consisting only of images of a shoe and of sunglasses. In completing task n, multi-task learning method cannot directly used on the support set for training, since they describe different concepts with different class labels. However, in order to perform well on any of the n tasks, the object must be recognized invariant of rotation, translation, lighting, scaling, among other features. Therefore, the images in the support set but not in the training set still provide valuable additional information to help increase generalization accuracy.

Several ways of modeling relationships across tasks have been explored and shown effective in experiments. Evgeniou et al. (2005) built a multi-task SVM classifier assuming all models were closed to each other and showed that this performed significantly better than its single-task counterpart. This proposed structure, where

the task parameters u_n were all close to an "average parameter" u_0, is known as regularized multi-task learning. In other applications, one can assume that the tasks share a common set of features (Argyriou et al., 2008). Other times, tasks may exhibit a more complicated group structure, known as clustered multi-task learning. Instead of each task being related to all others equally, tasks can be grouped such that they learn better from other tasks in the same group, and this group structure may be unknown a priori (Zhou et al., 2011).

In the recent study by Zhou et al. (2013), two novel multi-task learning methods are proposed to model AD progression, based on several previously established formulations of the multi-task relationships. Not only do they address the problem of prediction of scores at multiple future time points, which as mentioned above is far less studied in the literature, they also deal with the issue of high dimensional input data. For image data such as MRI, previously used methods of sequential feature evaluation (Ito et al 2010, Murphy 2010) are nonoptimal. Using PCA for dimension reduction could be effective, but would not result in a model that is interpretable, which is highly desirable given our goals of increasing our understanding of AD as well as seeking out ways to develop possible treatments. The formulations in Zhou et al. (2013) include penalty terms to effectively conduct feature selection in the training of the regression model, and are demonstrated to be effective using data from the first phase of the Alzheimer's Disease Neuroimaging Initiative (ADNI).

A new method for fitting high dimensional multivariate regression models, known as calibrated multivariate regression (CMR), was proposed by Liu et al. (2014). While traditional models have a form such as

$$\min_B \frac{1}{\sqrt{n}}||Y - XB||_F^2 + \lambda||B||_{1,p} \tag{1.1}$$

the CMR is the solution to

$$\arg\min_B ||Y - XB||_{2,1} + \lambda||B||_{1,p} \tag{1.2}$$

which has two distinct advantages: the tuning parameter selection is independent of the largest σ_k in $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ where Σ is the covariance matrix of the noise matrix Z; and the tasks are calibrated due to the $\ell_{2,1}$ loss function. A smoothed proximal gradient (SPG) algorithm to efficiently solve the CMR is proposed, and CMR is shown through experiments that it outperforms existing multivariate regression methods (Liu et al., 2014). Including this modified loss function in a multi-task learning model, then, may give improved performance as well.

In this study, we analyze the performance of the single-task ridge and Lasso methods as well as the multi-task temporal group Lasso and convex fused sparse group Lasso methods on two separate data sets: first on a data set similar to that used in Zhou et al. (2013) with the goal of replicating their results, and then on an expanded, up-to-date data set consisting of subjects in the currently ongoing ADNI study, to test whether the multi-task methods also display superior performance in cognitive score prediction on different data sets obtained through alternate settings. Because a biomarker must be validated in multiple clinical studies, using this second data set also helps validate the features previously found to be predictive of AD progression. Finally, we propose three novel multi-task learning methods for the prediction of AD progression at different time points, utilizing the modified loss function from the CMR model. We term formulations the calibrated temporal group Lasso and calibrated convex fused sparse group Lasso. We show how to efficiently solve the associated optimizations using an SPG algorithm, and compare their performance on the current ADNI study data with that of their non-calibrated counterparts.

Chapter 2

Data

All data in this study comes from the Alzheimers Disease Neuroimaging Initiative (ADNI) database, which can be accessed at `adni.loni.ucla.edu`. ADNI is a currently ongoing longitudinal study where a range of measurements are collected repeatedly every six or 12 months, with the goal of developing biomarkers in order to successfully detect and track Alzheimers disease (AD) in its early stages. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. ADNI began in 2004 with its initial phase, now known as ADNI1, which included 200 subjects diagnosed with early AD, 400 subjects diagnosed with mild cognitive impairment (MCI), and 200 cognitively normal (CN) elderly subjects serving as controls. AD participants meet the NINCDS/ADRDA criteria for probable AD, while MCI participants have reported memory concerns but no other significant levels of impairment or signs of dementia. CN participants show no signs of depression, cognitive impairment, or dementia. The ADNI1 study lasted for six years funded by the National Institute on Aging, 13 pharmaceutical companies, and two foundations that provided financial support through the Foundation for the National Institutes on Health.

ADNI GO, an extension of ADNI1, began in 2009 with 500 MCI and CN participants carried over from the first phase and 200 new participants, classified as having early mild cognitive impairment (EMCI). ADNI GO ended in 2011 as the currently ongoing phase, ADNI 2, began with another round of funding from its financial supporters. ADNI 2 consists of 150 new CN participants, between 450 to 500 CN and MCI participants from ADNI1, 150 new EMCI and 200 ADNI GO EMCI participants, 200 new AD participants, and 150 new participants identified as having late mild cognitive impairment (LMCI).

The MRI image features in this study came from the imaging data in the ADNI database processed by a team from University of California, San Francisco. Using the FreeSurfer image analysis suite, which can be found at `http://surfer.nmr.mgh.harvard.edu/`, they performed cortical reconstruction and volumetric segmentations to extract numeric feature values. In ADNI1, all participants received 1.5 Tesla (T) structural MRI. In ADNI2, all new participants are scanned using the 3T protocol, as are the CN, MCI, and EMCI participants carried forward from ADNI1 and ADNI GO. In order to do meaningful regression on input data, we want to limit extraneous noise in our feature values. To gauge whether the extracted feature values under the 3T protocol can be used in conjunction with the extracted feature values under the 1.5T protocol, we compared feature values from both at the same time points for subjects who had both forms of MRI scans. We found a high amount of variation between 1.5T and 3T scans, and determined the image data were not comparable and therefore separated the available ADNI MRI data into two groups, one consisting of baseline features from 1.5T scans of ADNI1 participants, and one consisting of baseline features from 3T scans of ADNI2 participants (including those carried over from ADNI1 and ADNI GO). We preprocessed both data sets by doing the following:

- features with more than 1000 missing entries (for all patients and all time points) were removed [1]
- image records that had a failed overall quality control score were removed
- patients without baseline MRI records were excluded
- the few missing entries were completed using the average value of the feature across all subjects

Zhou et al. 2013 Data Time Point	MMSE	ADAS-Cog
Baseline	648	648
M06	648	648
M12	642	638
M24	569	564
M36	389	377
M48	87	85
Data Set 1 (ADNI1) **Time Point**	**MMSE**	**ADAS-Cog**
Baseline	770	770
M06	770	769
M12	718	715
M24	630	627
M36	440	439
M48	110	109
Data Set 2 (ADNI2) **Time Point**	**MMSE**	**ADAS-Cog 13**
Baseline	1020	861
M06	881	785
M12	794	702
M24	622	538
M36	247	195
M48	80	70

Table 2.1: Summary of available ADNI data used in this study. The number of MRI features for each data set after preprocessing was 305, 327, and 340, respectively.

We also downloaded clinical assessment data from the ADNI database. In the ADNI1 phase, Mini Mental State Examination (MMSE) and Alzheimers Disease Assessment Scale cognitive subscale (ADAS-Cog) scores for all participants were collected at the baseline and at every follow-up visit. In the ADNI GO and ADNI2

[1]Specifically, these were features ST100SV, ST122SV, ST126SV, ST22CV, ST22SA, ST22TA, ST22TS, ST28CV, ST33SV, ST41SV, ST63SV, ST67SV, ST81CV, ST81SA, ST81TA, ST81TS, ST87CV, ST92SV and ST8SV.

phases, MMSE and ADAS-Cog 13, a modified version of the ADAS-Cog exam, was administered to all participants at every time point. The schedule for follow-up visits, however, differs between phases of the ADNI study. Because of this, and because the scores for ADAS-Cog 13 and ADAS-Cog are not comparable, for each assessment we again create two separate sets of target values, one for ADNI1 and one for the ongoing ADNI2. Two versions of ADAS-Cog total scores, TOTAL11 and TOTALMOD, were available. In this study we utilize the former, which had fewer missing values for participants in the ADNI1 phase. A summary of the data described above is given in Table 2.1. Note that the clinical scores for many patients are missing at certain time points, due to a failure to collect data or to participants dropping out of the study entirely. We could not recreate the exact data set used in Zhou et al. (2013) - the original preprocessed data set has since been misplaced by the authors since 2011, and the ADNI database appears to have restructured its available data since that time.

Chapter 3

Methodology and Models

The research was divided into three parts. In the first, we are concerned only with replication of Zhou et al. (2013). We apply four models, two single task and two multi-task whose formulation and implementation algorithms were proposed in the paper. In the second phase, we expand the dataset to include all current subjects in the ADNI1, ADNI GO, and ADNI2 studies, which also included additional time points of target value collection. The same four models were applied and the results compared to the subset of data from the first phase. In the final phase, we propose a new model with a modified objective function, based on the method of calibrated multivariate regression (Liu et al., 2014), and compare its performance against the earlier established methods.

Throughout this study, we use as an indicator of cognitive function the cognitive scores of patients at multiple time points. We treat the prediction of cognitive scores at a given time point as a regression problem, and thus formulate the prediction of cognitive scores at multiple future time points as the multi-task version of the regression problem. This is made possible by the assumption that cognitive scores between subsequent time points should be smooth and therefore the prediction tasks are related. Our goal is to solve the multi-task regression problem of predicting cog-

nitive scores for t time points given d baseline MRI features of n subjects. We utilize linear models for the prediction: given $x_i \in R^d$, the baseline features for patient i, and $w^s \in R^d$, the weight vector for the features at time point s, the predicted target value (clinical score) for patient i at time point s is $x_i^T w^s$. Our models employ data matrix $X = [x_1, \cdots, x_n]^T \in R^{nxd}$, target matrix $Y = [y_1, \cdots, y_t]^T \in R^{nxt}$, and weight matrix $W = [w^1, \cdots, w^t] \in R^{dxt}$. Note that due to missing cognitive assessment data, our matrix Y is incomplete.

Ridge

In typical least squares regression, a coefficient or weight matrix W is found by minimizing the empirical error on the training data:

$$\min_W ||XW - Y||_F^2 \tag{3.1}$$

where $||W||_F^2 = \sum_{i=1}^{d} \sum_{j=1}^{t} W_{i,j}^2$. This yields a solution

$$\hat{W} = (X^T X)^{-1} X^T Y \tag{3.2}$$

which, while it is simple and produces a weight matrix that gives the least amount of error on the training data, often gives a higher error on the testing data, due to overfitting on the training data. By adding a regularization term to the objective function, a penalty is imposed on the size of the coefficients. In ridge regression, the problem becomes

$$\min_W ||XW - Y||_F^2 + \lambda ||W||_F^2 \tag{3.3}$$

Given the nature of the MRI data we are working with, namely the reasonable assumptions that features of the brain are highly correlated and that a small subset of the features is predictive of the progression of AD, ridge regression is expected to improve the prediction performance since coefficients are shrunk toward 0 and therefore

each other. Rewriting the above formulation as

$$\min_W (XW - Y)^T (XW - Y) + \lambda W^T W \tag{3.4}$$

taking the derivative with respect to W we get the analytical solution

$$\hat{W}^{ridge} = (X^T X + \lambda I)^{-1} X^T Y \tag{3.5}$$

Lasso

Like ridge regression, Lasso (Tibshirani, 1996) shrinks the coefficients, only with a different regularization term. It solves the optimization problem

$$\min_W ||XW - Y||_F^2 + \lambda ||W||_1 \tag{3.6}$$

where $||W||_1 = \sum_{i=1}^{d} \sum_{j=1}^{t} |W_{i,j}|$. Due to this term, the solutions are no longer linear on the entries of Y, which means there is no closed-form expression for the optimal weight matrix. Another difference is that with this term, shrinkage results in some of the optimal coefficients equaling zero. As a result, we get a sparse coefficient matrix, which translates into more interpretable results. For a geometric illustration of the properties of ridge and Lasso estimators, see Figure 3.1.

To solve for $\hat{W}^{lasso}$, which because of its non-smooth term does not have an analytical solution like ridge regression, we implement the `glmnet` package in R, which uses the extremely efficient method of cyclical coordinate descent to fit an elastic-net regularization path for linear regression (Friedman et al., 2010).

As mentioned previously, our true target value matrix Y is incomplete. Instead of solving the textbook ridge and Lasso formulations described above, our true optimization problems then become

$$\min_W ||S \odot (XW - Y)||_F^2 + \lambda ||W||_F^2 \tag{3.7}$$

and

$$\min_W ||S \odot (XW - Y)||_F^2 + \lambda ||W||_1 \tag{3.8}$$

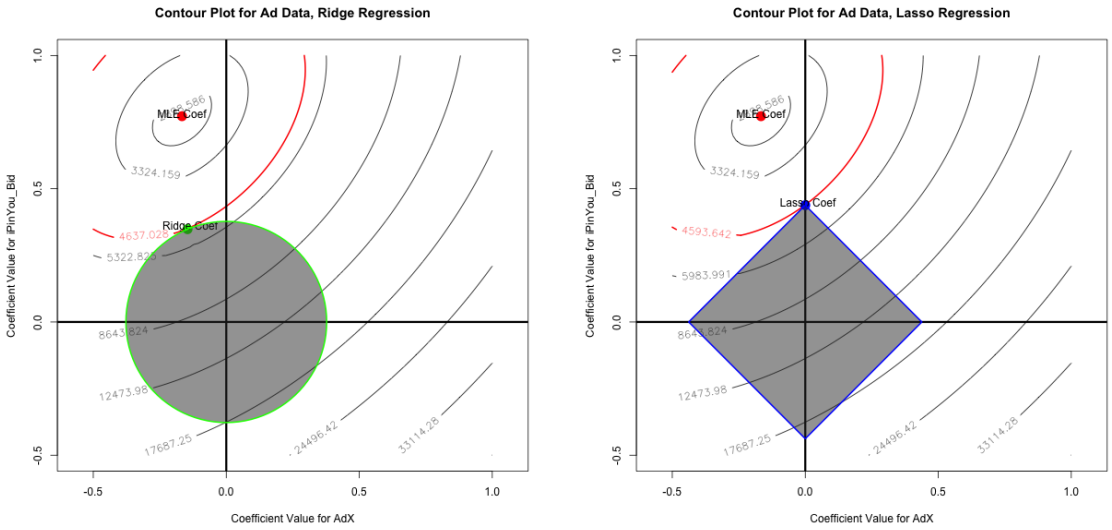

Figure 3.1: Ridge vs lasso regression, illustrated for the case $d = 2$. Because ridge regression uses the ℓ_2 norm, rewriting the optimization gives us an upper bound on the coefficients restricting them to lie within an area in the shape of a circle. On the other hand, the ℓ_1 norm, which is the sum of the absolute value of the coefficients, causes the intersection between the level curves of the MLE and the coefficient boundary to be on a point, resulting in sparse coefficient matrices.

where $S \in \mathbb{R}^{nxt}, S_{i,j} = 0$ if the score for subject i at time point j is missing, and $S_{i,j} = 1$ otherwise. $Z = A \odot B$ denotes $Z_{i,j} = A_{i,j} * B_{i,j}$ for all i and j. However, because the missing value matrix S only shows up in the first term, its only effect is to remove the contribution to empirical error by the prediction of the score for patient i at time j. Therefore, since both the ridge and Lasso models are single task models, we can build W column-by-column in a vectorized manner. That is, because

$$W = [w^1, \cdots, w^t] = (X^T X + \lambda I)^{-1} X^T [Y_1 \cdots Y_t] \tag{3.9}$$

at each time point i we can find the optimal coefficients for the d features given input matrix $X_i \in \mathbb{R}^{n_i xd}$ and $Y_i \in \mathbb{R}^{n_i}$, where n_i is the number of actually available assessment scores at time i, so that

$$w^i = (X_i^T X_i + \lambda I)^{-1} X_i^T Y_i \tag{3.10}$$

Decomposing the matrix problem into vector form in this way assumes the tasks are independent of each other, which in reality is not the case, since the ADNI study is longitudinal and tasks are related to each other in time.

Temporal Group Lasso

When studying the onset and progression of disease over time, numeric values representing the state of the illness are sometimes assumed to be monotonically changing or to have little difference between successive time points. Since we use linear models for prediction in this study ($\hat{y}_i^s = x_i^T w^s$), the difference between predictions for the same patient between consecutive time points can be written

$$|\hat{y}_i^{s+1} - \hat{y}_i^s| = |x_i^T w^{s+1} - x_i^T w^s| = |x_i^T (w^{s+1} - w^s)| \tag{3.11}$$

If we want a model that predicts cognitive assessment scores that are more smooth temporally, we can discourage a large $|\hat{y}_i^{s+1} - \hat{y}_i^s|$ by penalizing large deviations between the models at successive time points. We add a regularization term to the objective function of the ridge regression formulation to get:

$$\min_W ||S \odot (XW - Y)||_F^2 + \lambda_1 ||W||_F^2 + \lambda_2 \sum_{i=1}^{t-1} \|w^{i+1} - w^i\|_2^2 \tag{3.12}$$

If we define $H \in \mathbb{R}^{tx(t-1)}$ as $H_{i,j} = 1$ if $i = j, H_{i,j} = -1$ if $i = j + 1$, and $H_{i,j} = 0$ otherwise, then the above can be rewritten as

$$\min_W ||S \odot (XW - Y)||_F^2 + \lambda_1 ||W||_F^2 + \lambda_2 \|WH\|_F^2 \tag{3.13}$$

Additionally, we have to deal with the problem of having a large number of features and a low number of subjects for whom we have target values, especially for later time points (see Table 2.1). Our prediction model as it is suffers from the "curse of dimensionality". Principal component analysis (PCA), which takes the original set of observed variables and transforms it into a space based on sets of uncorrelated variables, would help reduce the dimensionality of our input data. However, because the components in PCA are linear combinations of the original features, the model we get after fitting a weight matrix to these components will be difficult to interpret. Other traditional methods of feature selection are not appropriate for our multi-task regression here, since every subject has a different pattern of missing target values at each task.

Instead, we add a group Lasso regularization term based on the $\ell_{1,2}$ penalty (Yuan and Lin, 2006). Similar to the effect the regular ℓ_1 norm has in Lasso regression, this penalty results in sparseness and therefore selects a small set of features, but does so such that the set of features is the same across tasks. Adding this to the objective function above, we thereby obtain the temporal group Lasso formulation:

$$\min_W ||S \odot (XW - Y)||_F^2 + \lambda_1 ||W||_F^2 + \lambda_2 \|WH\|_F^2 + \lambda_3 \|W\|_{1,2} \tag{3.14}$$

where $||W||_{1,2} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{t} W_{i,j}^2}$. The square-root term within the outer sum is the ℓ_2 norm of the weights of a particular feature over all tasks, and the outer ℓ_1 norm thus effectively selects features based on how strong of a predictor it is over all tasks. To solve for this optimization problem, we utilize the accelerated gradient method (AGM) involving the computation of the proximal operator associated with the group Lasso penalty. We employ the algorithm used in the MALSAR package, which does this optimization efficiently (Zhou et al.).

Convex Fused Sparse Group Lasso

While the TGL formulation utilizes a strong notion of multi-task learning, that is, restricting the models at all time points to share a fixed common set of features, we would like to incorporate more sophistication into how the model learns from multiple tasks. Previous studies (Thompson et al., 2004; Caroli and Frisoni, 2010) have shown that biomarkers observed in MRI differ in their patterns over time during the progression of AD. Therefore, we want to allow our model to select features in a way that still limits how many are predictive of the progression, but these sets are allowed to change slightly over time. That is, a feature such as volume of grey matter in a region of the hippocampus may be included in the a set predictive of scores at the first few time points, but later as the disease progresses, the loss rate may decline while ventricles expand (Thompson et al., 2004), and so we would no longer include that

feature for score prediction in the future. To achieve this, we utilize regularization terms that balance joint feature selection for multiple tasks with task-specific feature selection, and maintaining temporal smoothness. From these requirements we get the convex fused sparse group Lasso (cFSGL) formulation (Zhou et al., 2013):

$$\min_{W} ||S \odot (XW - Y)||_F^2 + \lambda_1 ||W||_1 + \lambda_2 \|WH\|_1 + \lambda_3 \|W\|_{1,2} \tag{3.15}$$

where the combination of $||W||_1$, the Lasso penalty, and $||W||_{1,2}$, the group Lasso penalty from above, is known as the sparse group Lasso penalty. These two in combination allow for simultaneous selection of specific features for each task and joint feature selection for all tasks. Temporal smoothness is incorporated in the $||WH||_1$ term, called the fused Lasso penalty.

This optimization problem is even more difficult to solve than the TGL formulation, since it involves three non-smooth terms. Again, this can be solved using AGM, which involves the computation of the proximal operator associated with the cFSGL formulation. What makes this one harder is that while the computation for the proximal operator for the $\ell_{1,2}$ has been established for a while, the computation for the proximal operator with three non-smooth terms relies on a decomposition property to be efficiently computed (Zhou et al., 2013). We again employ the algorithm used in the MALSAR package. For a comparison of Lasso, TGL and cFSGL models, see Figure 3.2

Calibrated Multi-task

Since calibrated multivariate regression has been shown to outperform existing multivariate regression in certain cases (Liu et al., 2014), we would like to extend the formulation to the multi-task problem of predicting cognitive assessment scores at multiple time points.

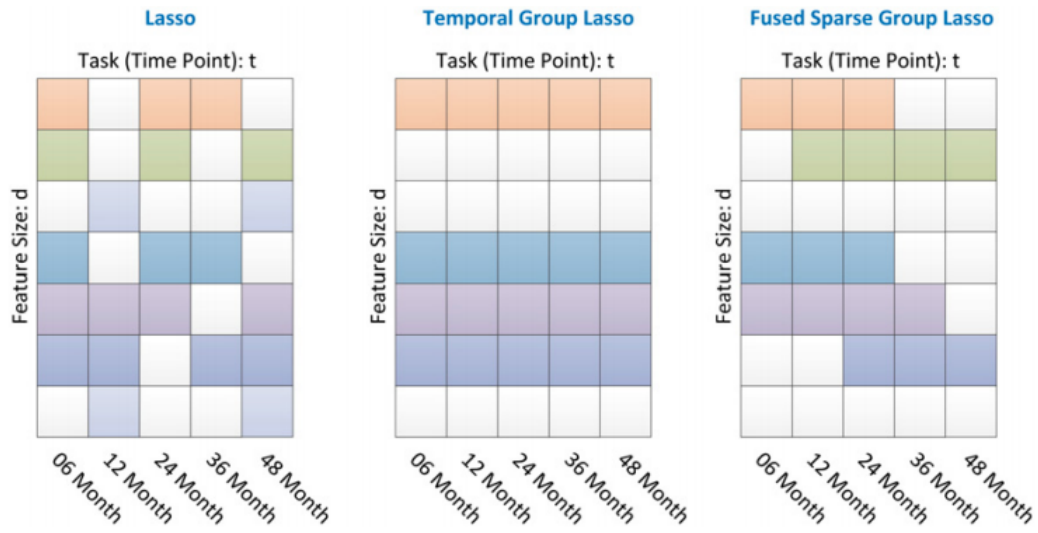

Figure 3.2: A comparison of the models built by different approaches (Zhou et al 2013). In Lasso, which is a single-task learning method, the features selected at each time point are done so independently, so no general sparsity pattern is observed. In TGL, the models for all time points must share the same set of features. In cFSGL, the set of selected features across time points are not identical but are smooth. M12 features differ from M06 feature by one, M24 from M12 by one, M36 from M24 by two, and M48 from M36 by one.

The first model we develop is the calibrated ridge with group Lasso penalty:

$$\min_{W} ||S \odot (XW - Y)||_{2,1} + \lambda_1 ||W||_F^2 + \lambda_2 \|W\|_{1,2} \tag{3.16}$$

which can be solved using the dual smoothing technique given in Han et al. That is, by considering the Fenchel's dual representation of the $\ell_{2,1}$ loss, we can rewrite:

$$||XW - Y||_{2,1} = \max_{||U||_{2,\infty} \leq 1} \langle U, XW - Y \rangle \tag{3.17}$$

Using a smoothing parameter μ, the smooth approximation of the above can be obtained through solving

$$||XW - Y||_{\mu} = \max_{||U||_{2,\infty} \leq 1} \langle U, XW - Y \rangle - \frac{\mu}{2} ||U||_F^2 \tag{3.18}$$

which has the closed form solution $\hat{U}^W$, where

$$\hat{U}^W_{*k} = (XW_{*k} - Y_{*k}) / \max\{||XW_{*k} - Y_{*k}||_2, \mu\} \tag{3.19}$$

Han et al showed this approximation $||XW - Y||_{\mu}$ to be smooth in W, and $G^{\mu}(W)$, its gradient with respect to W, equals $X^T \hat{U}^W$. This smoothed version of our original objective function, $\min_{W} ||XW - Y||_{\mu} + \lambda_1 ||W||_F^2 + \lambda_2 \|W\|_{1,2}$, can then be solved by

adopting the fast proximal gradient algorithm (Liu et al., 2014). The code to do so for this specific formulation can be found in the MALSAR package (Zhou et al.).

In our formulation, where we allow missing target values in Y, we incorporate the S matrix simply by a decomposition of the $\ell_{2,1}$ norm, replacing instances of $||XW - Y||_{2,1}$ with $\sum_{i=1}^{t} ||X_i W_i - Y_i||_2$, where once again $X_i \in \mathbb{R}^{n_i x d}$ and $Y_i \in \mathbb{R}^{n_i}$, where n_i is the number of actually available assessment scores at time i, and $W_i \in \mathbb{R}^d$ is the ith column of the weight matrix.

We want to incorporate the desired temporal smoothness of the target scores into our calibrated model formulation. For this we have the calibrated TGL model which solves the optimization

$$\min_W ||S \odot (XW - Y)||_{2,1} + \lambda_1 ||W||_{1,2} + \lambda_2 \|W\|_F^2 + \lambda_3 ||WH||_F^2 \tag{3.20}$$

The only difference between the calibrated TGL and the calibrated ridge with group Lasso is the addition of the temporal smoothness term, which itself is smooth and easily differentiable with respect to W. Minor adjustments to the fast proximal gradient algorithm used above then gives us an efficient way to compute the optimal W. [1]

Finally, since the cFSGL formulation has been shown to outperform the TGL in some cases, we additionally propose the calibrated cFSGL formulation

$$\min_W ||S \odot (XW - Y)||_{2,1} + \lambda_1 \|W\|_{1,2} + \lambda_2 ||W||_1 + \lambda_3 ||WH||_1 \tag{3.21}$$

which is more difficult to solve than the calibrated TGL due to its three non-smooth regularization terms. To reduce the optimization problem to be more similar to the convex program from Liu et al. (2014), we rewrite the above as

$$\min_W ||S \odot (XW - Y)||_{2,1} + \lambda_1 \|W\|_{1,2} + ||W\tilde{H}||_1 \tag{3.22}$$

[1] See Appendix for MATLAB and C code written for this computation.

where $\tilde{H} = [\lambda_2 I, \lambda_3 H]$. We then apply the same dual smoothing technique used on the loss function, rewriting

$$||W\tilde{H}||_1 = \max_{||V||_\infty \leq 1} \text{Tr}(W\tilde{H}V^T) \tag{3.23}$$

to get the smooth approximation

$$||W\tilde{H}||_\mu = \max_{||V||_\infty \leq 1} \text{Tr}(W\tilde{H}V^T) - \frac{\mu}{2}||V||_F^2 \tag{3.24}$$

Note, however, that this objective function is separable, since $\text{Tr}(W\tilde{H}V^T) = \sum_{i,j}(W\tilde{H})_{ij}V_{ij}$ and $||V||_F^2 = \sum_{i,j} V_{ij}^2$. Then the optimization problem becomes

$$\max_{|V_{i,j}| \leq 1} (W\tilde{H})_{ij}V_{ij} - \frac{\mu}{2}V_{ij}^2 \tag{3.25}$$

which we can easily see is quadratic in V_{ij}, allowing us to solve for $\hat{V}^W$ elementwise:

$$\hat{V}_{ij} = \begin{cases} \frac{(W\tilde{H})_{ij}}{\mu} & \text{if } \frac{(W\tilde{H})_{ij}}{\mu} \in [-1, 1] \\ 1 & \text{if } \frac{(W\tilde{H})_{ij}}{\mu} > 1 \\ -1 & \text{if } \frac{(W\tilde{H})_{ij}}{\mu} < -1 \end{cases} \tag{3.26}$$

The approximation $||W\tilde{H}||_\mu$ is smooth in W and its gradient $G^\mu(W) = \hat{V}^W\tilde{H}^T$. Further adjustments to the proximal gradient algorithm to incorporate this smoothed approximation allows us to solve for the optimal W in this formulation. [2]

[2]See Appendix for MATLAB and C code written for this computation.

Chapter 4

Results

In this section, we test the prediction performance of our models built on different data sets in each part of this study. Throughout, we use three measures of overall regression performance: root mean square error (rMSE), normalized mean square error (nMSE), and weighted correlation coefficient (wR), defined as follows:

$$\text{rMSE}(Y_i, \hat{Y}_i) = \frac{||Y_i - \hat{Y}_i||_2}{\sqrt{n_i}} \tag{4.1}$$

$$\text{nMSE}(Y, \hat{Y}) = \frac{\sum_{i=1}^{t} ||Y_i - \hat{Y}_i||_2^2 / \sigma^2(Y_i)}{\sum_{i=1}^{t} n_i} \tag{4.2}$$

$$\text{wR}(Y, \hat{Y}) = \frac{\sum_{i=1}^{t} \text{Corr}(Y_i, \hat{Y}_i) * n_i}{\sum_{i=1}^{t} n_i} \tag{4.3}$$

where n_i is the number of available true target values at time point i, $Y_i \in \mathbb{R}^{n_i x 1}$ is the vector of available ground truth target values, $\hat{Y}_i$ is the vector of corresponding predicted values, and Corr is the correlation coefficient between two vectors (Zhou et al., 2013). While the root mean square error measures the error within each time point prediction, normalized mean square error has been used in multi-task learning literature (Argyriou et al., 2008) and the weighted R-value has been employed in previous literature on AD progression.

Table 4.1: Comparison of single-task ridge and Lasso with multi-task temporal group Lasso and convex fused sparse group Lasso on subset of ADNI1 data reported in Zhou et al. (2013). Tasks are the prediction of MMSE and ADAS-Cog scores at five future time points using only baseline MRI and baseline MMSE.

	Ridge	Lasso	TGL	cFSGL
MMSE:				
nMSE	0.548 ± 0.057	0.459 ± 0.042	0.449 ± 0.045	**0.395 ± 0.052**
wR	0.689 ± 0.030	0.746 ± 0.031	0.755 ± 0.029	**0.796 ± 0.031**
M06 rMSE	2.269 ± 0.207	2.071 ± 0.261	**2.038 ± 0.262**	2.071 ± 0.213
M12 rMSE	3.266 ± 0.556	2.973 ± 0.654	2.923 ± 0.643	**2.762 ± 0.669**
M24 rMSE	3.494 ± 0.599	3.371 ± 0.747	3.363 ± 0.733	**3.000 ± 0.642**
M36 rMSE	4.003 ± 0.853	3.786 ± 0.926	3.768 ± 0.962	**3.265 ± 0.803**
M48 rMSE	4.328 ± 1.310	3.653 ± 1.268	3.631 ± 1.226	**2.871 ± 0.884**
ADAS-Cog:				
nMSE	0.532 ± 0.095	0.520 ± 0.084	0.464 ± 0.067	**0.391 ± 0.059**
wR	0.705 ± 0.043	0.716 ± 0.036	0.747 ± 0.033	**0.803 ± 0.024**
M06 rMSE	5.213 ± 0.522	4.976 ± 0.518	4.820 ± 0.489	**4.451 ± 0.340**
M12 rMSE	6.079 ± 0.775	6.193 ± 0.766	5.813 ± 0.697	**5.230 ± 0.589**
M24 rMSE	7.409 ± 1.154	7.275 ± 1.099	6.835 ± 1.052	**6.249 ± 0.996**
M36 rMSE	7.143 ± 1.351	7.139 ± 1.444	6.938 ± 1.363	**5.928 ± 1.064**
M48 rMSE	6.644 ± 2.750	6.879 ± 2.465	6.000 ± 2.738	**5.980 ± 1.979**

Part I: ADNI1 Study Replication

In this section, we compare our four different modeling approaches for predicting future MMSE and ADAS-Cog scores of the ADNI1 cohort, using baseline MRI images and baseline MMSE scores as input. We independently apply the models on MMSE and ADAS-Cog target values and do not assume the scores are correlated. We report the mean and standard deviation of the prediction performance measures, based on 20 iterations of different random splits of data where 90% is used as training data, and compare them to the predictive performance reported in existing studies in Table 4.1. For ridge and Lasso, we used 5-fold cross validation to fit the model parameter on the training data. For TGL and cFSGL, we chose model parameters that, among 20 random splits of data with a 9:1 training to testing ratio, gave the lowest average nMSE on the test sets, and reported the averaged performance measures for those splits with those parameters. Experimental results are presented in Table 4.2

Table 4.2: Comparison of single-task ridge and Lasso with multi-task temporal group Lasso and convex fused sparse group Lasso on ADNI1 data. Tasks are the prediction of MMSE and ADAS-Cog scores at five future time points using only baseline MRI and baseline MMSE.

	Ridge	Lasso	TGL	cFSGL
MMSE:				
nMSE	0.5402 ± 0.0642	0.4687 ± 0.0627	**0.4463 ± 0.0529**	0.4607 ± 0.0613
wR	0.6925 ± 0.0458	0.7527 ± 0.0442	**0.7663 ± 0.0418**	0.7562 ± 0.0479
M06 rMSE	2.4905 ± 0.2312	2.2688 ± 0.2505	**2.2120 ± 0.2551**	2.2623 ± 0.2522
M12 rMSE	2.7739 ± 0.3028	2.5305 ± 0.2934	**2.4785 ± 0.3563**	2.5027 ± 0.2709
M24 rMSE	3.3853 ± 0.4044	3.2088 ± 0.4278	**3.2649 ± 0.5731**	3.2676 ± 0.4906
M36 rMSE	3.9124 ± 0.6205	3.7638 ± 0.7081	3.6942 ± 0.9048	**3.6416 ± 0.7462**
M48 rMSE	3.7755 ± 1.1174	3.5310 ± 1.1186	3.6450 ± 1.3377	**3.5983 ± 1.1734**
ADAS-Cog:				
nMSE	0.5433 ± 0.0757	0.4920 ± 0.0582	**0.4801 ± 0.0553**	0.4819 ± 0.0544
wR	0.6901 ± 0.0553	0.7388 ± 0.0442	0.7456 ± 0.0474	**0.7459 ± 0.0450**
M06 rMSE	5.2119 ± 0.7866	4.8981 ± 0.8309	**4.7081 ± 0.9237**	4.8468 ± 0.8353
M12 rMSE	5.8140 ± 0.7362	5.4722 ± 0.6105	5.6029 ± 0.7573	**5.4306 ± 0.6360**
M24 rMSE	6.9142 ± 0.6330	6.5825 ± 0.8329	6.6279 ± 1.0070	**6.5763 ± 0.8549**
M36 rMSE	8.1648 ± 1.2620	8.1521 ± 1.4404	8.1692 ± 1.6426	**7.8919 ± 1.4704**
M48 rMSE	7.4674 ± 2.0543	6.8573 ± 2.1648	**7.0759 ± 2.5386**	7.0799 ± 2.1146

We found that on the currently available ADNI1 data, the multi-task learning models outperformed the single-task ones, as was the case in Zhou et al 2013. However, whereas previously it was found that cFSGL performed better than TGL on both MMSE and ADAS-Cog prediction, in this study relative performance of the two models differed depending on the prediction tasks. For the MMSE prediction, TGL performs better than cFSGL in terms of both nMSE and weighted correlation coefficient. For the ADAS-Cog prediction, overall performance of TGL and cFSGL were almost identical in terms of nMSE and weighted correlation coefficient. For both MMSE and ADAS-Cog, however, cFSGL performed essentially equal to or better than TGL for prediction at later time points.

Table 4.3: Comparison of single-task ridge and Lasso with multi-task temporal group Lasso and convex fused sparse group Lasso on ongoing ADNI2 data. Tasks are the prediction of MMSE and ADAS-Cog 13 scores at five future time points using only baseline MRI and baseline MMSE.

	Ridge	Lasso	TGL	cFSGL
MMSE:				
nMSE	0.4986 ± 0.0458	0.4548 ± 0.0647	0.4196 ± 0.0459	**0.4177 ± 0.0545**
wR	0.7170 ± 0.0355	0.7600 ± 0.0317	0.7765 ± 0.0341	**0.7805 ± 0.0350**
M06 rMSE	2.2017 ± 0.3283	2.0550 ± 0.2964	2.0250 ± 0.2994	**2.0203 ± 0.2925**
M12 rMSE	2.3965 ± 0.3675	2.2409 ± 0.3621	2.1948 ± 0.3570	**2.1805 ± 0.3528**
M24 rMSE	2.6767 ± 0.3439	2.4445 ± 0.2631	2.4282 ± 0.2981	**2.3986 ± 0.2857**
M36 rMSE	3.5623 ± 1.2111	3.4506 ± 1.1918	3.2964 ± 1.1445	**3.2347 ± 1.0714**
M48 rMSE	2.9535 ± 1.4864	3.0821 ± 1.2996	**2.8240 ± 1.3694**	2.8627 ± 1.2692
ADAS-Cog 13:				
nMSE	0.5096 ± 0.0850	0.4907 ± 0.1014	0.4609 ± 0.0770	**0.4603 ± 0.0889**
wR	0.7280 ± 0.0515	0.7436 ± 0.0501	0.7545 ± 0.0513	**0.7602 ± 0.0519**
M06 rMSE	6.5743 ± 0.6019	6.4339 ± 0.5323	6.3494 ± 0.5435	**6.3314 ± 0.5290**
M12 rMSE	7.3508 ± 1.2590	7.0406 ± 1.0930	6.9721 ± 1.1938	**6.9458 ± 1.1627**
M24 rMSE	8.3349 ± 1.4683	7.9888 ± 1.1824	7.9226 ± 1.2802	**7.8289 ± 1.1631**
M36 rMSE	8.0739 ± 2.1248	8.2968 ± 2.0953	**7.9794 ± 2.2666**	8.0030 ± 2.2163
M48 rMSE	8.7433 ± 4.7172	8.7578 ± 4.1977	8.5650 ± 4.7273	**8.4829 ± 4.5885**

Part II: Expansion to Ongoing ADNI Study Data

In this next section, we compare the same four modeling approaches for predicting future MMSE and, this time, ADAS-Cog 13 scores of the ADNI2 cohort, using baseline MRI features and baseline MMSE scores as input. We again apply the models on MMSE and ADAS-Cog 13 target values independently. We report the mean and standard deviation of the prediction performance measures, based on 20 iterations of different random splits of data where 90% is used as training data in Table 4.3. For ridge and Lasso, we used 5-fold cross validation to fit the model parameter on the training data. For TGL and cFSGL, we chose model parameters that, among 20 random splits of data with a 9:1 training to testing ratio, gave the lowest average nMSE on the test sets, and reported the average performance measures using those parameters for those same splits.

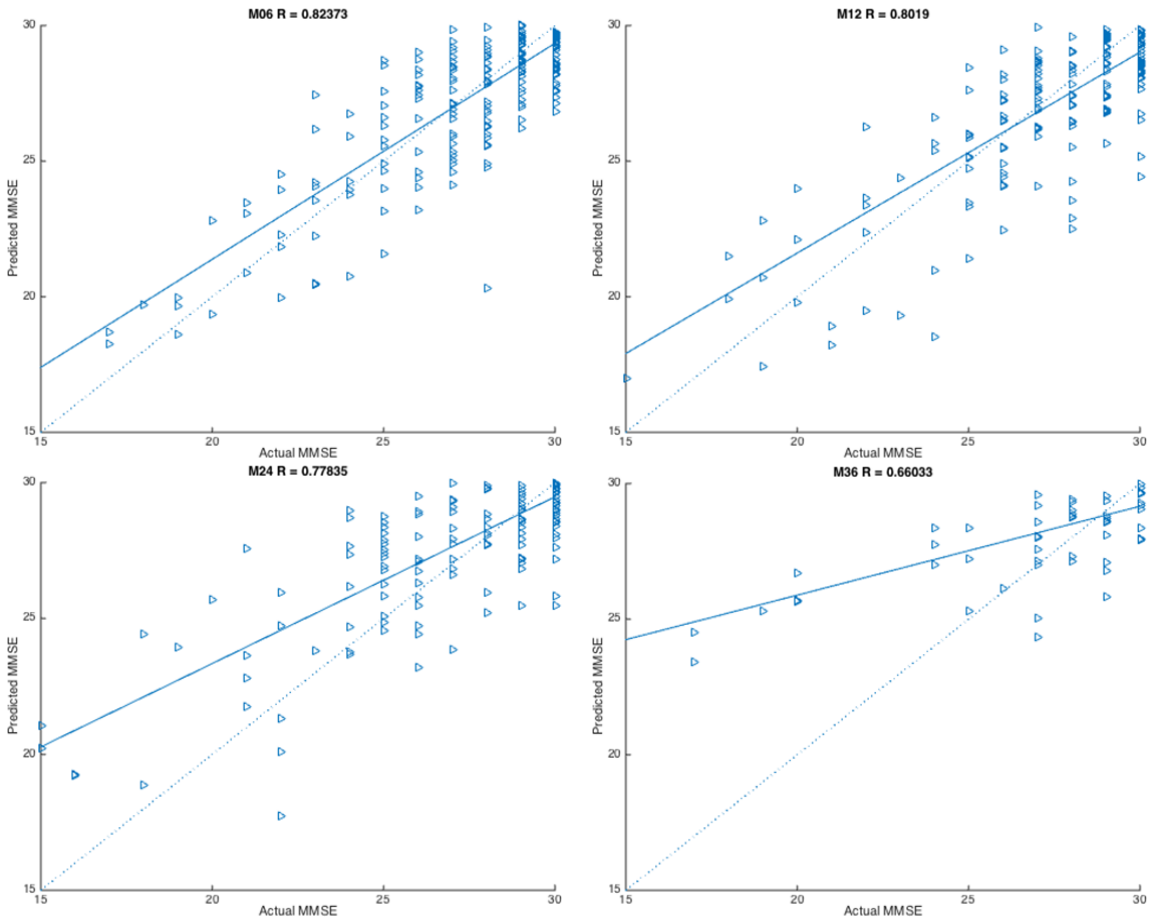

Figure 4.1: Scatter plots of actual versus predicted MMSE scores on testing data, from a model built by cFSGL using baseline MRI features and baseline MMSE as input. The dashed line represents perfect correlation, meaning the predicted score exactly matched the actual score. The solid line is the regression line from performing least squares on the points.

Again, for this updated data set, the multi-task learning models outperformed the single-task ones. cFSGL appears to perform better than TGL across the board, as Zhou et al. (2013) observed, however it did not do so to the same degree. That is, the difference between TGL and cFSGL rMSE values were not as large as observed in the study on an older data set. The improvement in weighted correlation coefficient from TGL to cFSGL was a little greater than in other measures, but again overall performance is remarkably similar. To visualize prediction performance for each time point, we show the scatter plots for predicted versus actual scores, for both MMSE and ADAS-Cog 13, in Figure 4.1 and Figure 4.2 respectively. We only show the first four time points since M48 had very few available samples especially in the testing

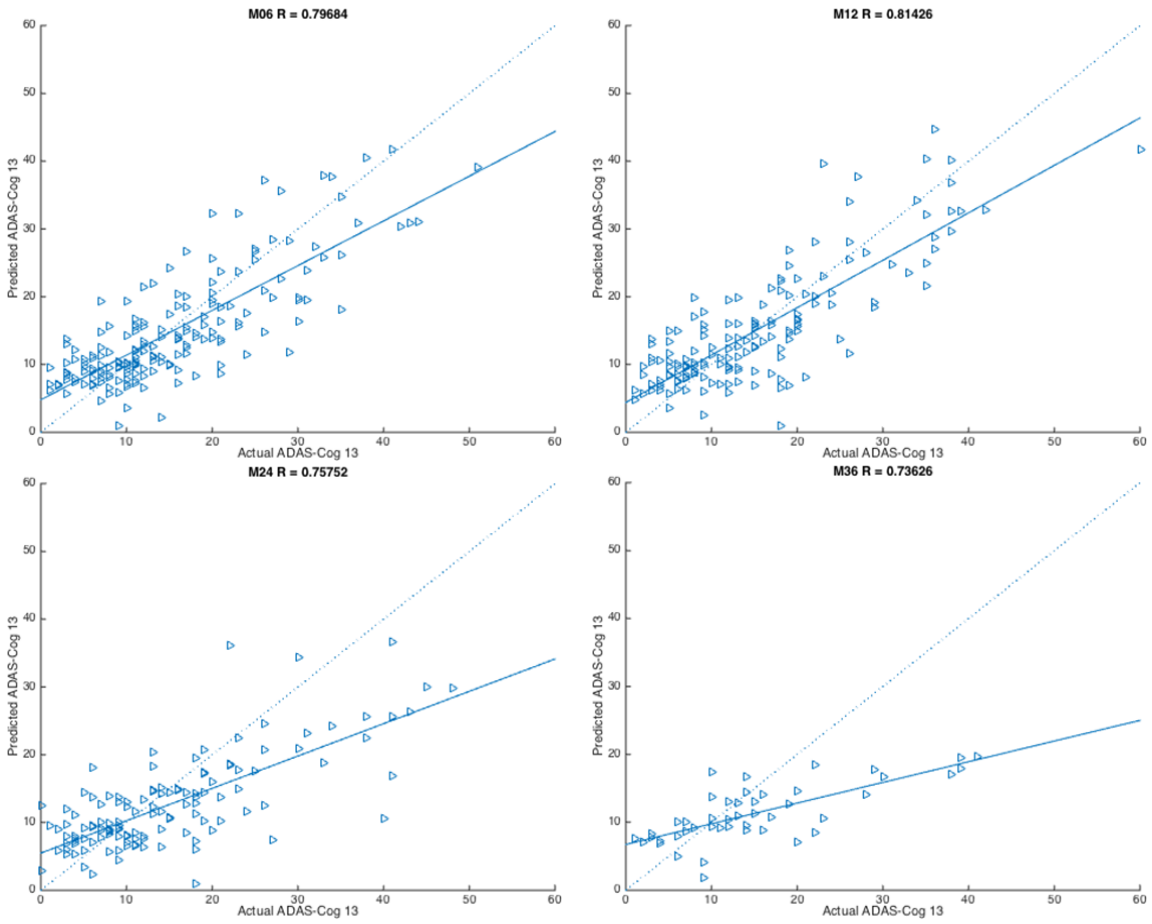

Figure 4.2: Scatter plots of actual versus predicted ADAS-Cog 13 scores on testing data, from a model built by cFSGL using baseline MRI features and baseline MMSE as input. The dashed line represents perfect correlation, meaning the predicted score exactly matched the actual score. The solid line is the regression line from performing least squares on the points.

set. For illustrative purposes, we display the test results on 25% of the data from training on 75% of the original data, using the same model parameters that resulted in the average nMSE values in Table 4.3.

Part III: A Novel Approach

In the final part of our study, we build a prediction model again using baseline MRI features and baseline MMSE from the ADNI2 data set, using the calibrated multi-task formulations (calibrated ridge with group Lasso, calibrated TGL, and calibrated cFSGL) sproposed earlier. Models are applied on the target values indepen-

Table 4.4: Performance of our three novel calibrated multi-task regression models, with different regularizations, on ADNI2 data. Tasks are the prediction of MMSE and ADAS-Cog 13 scores at five future time points using only baseline MRI and baseline MMSE.

	Calibrated Ridge + $\ell_{1,2}$	Calibrated TGL	Calibrated cFSGL
MMSE:			
nMSE	0.4282 ± 0.0448	0.4217 ± 0.0481	**0.4184 ± 0.0464**
wR	0.7686 ± 0.0327	0.7777 ± 0.0337	**0.7786 ± 0.0332**
M06 rMSE	2.0387 ± 0.3087	2.0231 ± 0.3064	**2.0172 ± 0.2986**
M12 rMSE	2.2047 ± 0.3587	2.2004 ± 0.3648	**2.1970 ± 0.3525**
M24 rMSE	2.4358 ± 0.2810	2.4290 ± 0.3035	**2.4187 ± 0.2928**
M36 rMSE	3.3899 ± 1.2155	**3.2784 ± 1.1068**	3.3096 ± 1.1597
M48 rMSE	2.8027 ± 1.3387	2.8186 ± 1.2800	**2.7639 ± 1.3460**
ADAS-Cog 13:			
nMSE	0.4751 ± 0.0827	0.4668 ± 0.0820	**0.4599 ± 0.0781**
wR	0.7497 ± 0.0502	0.7526 ± 0.0541	**0.7561 ± 0.0521**
M06 rMSE	6.3988 ± 0.5835	6.3905 ± 0.5727	**6.3398 ± 0.5367**
M12 rMSE	7.0236 ± 1.2022	7.0243 ± 1.2206	**6.9746 ± 1.1747**
M24 rMSE	7.9885 ± 1.3485	7.9330 ± 1.2748	**7.8778 ± 1.2375**
M36 rMSE	7.9817 ± 2.1959	8.0386 ± 2.2558	**7.9723 ± 2.2846**
M48 rMSE	8.5184 ± 4.4961	8.5178 ± 4.6622	**8.4528 ± 4.6853**

dently. We report the mean and standard deviation of the prediction performance measures, based on 20 random splits of data into training and testing sets using a ratio of 9:1, in Table 4.4. For all three, we chose model parameters that, among these 20 splits, gave the lowest average nMSE on the test sets, and reported the average performance measures using those parameters for those same splits.

Given the earlier results from Part II, we would expect the calibrated cFSGL to outperform the calibrated TGL, and the results above show that this is indeed the case. Calibrated TGL also appears to slightly outperform calibrated ridge with group Lasso, indicating that the temporal smoothness term improves the predictive power of our resulting model. However, based on the advantages of calibrated multivariate regression over ordinary multivariate regression, we had expected the calibrated multi-task formulations to perform better than their non-calibrated counterparts. Instead, they seem to do slightly worse on the same 20 splits of data, when using model parameters fit to minimize average nMSE on those same 20 splits. We fix these

Table 4.5: Comparison of MMSE prediction from calibrated versus non-calibrated TGL on 20 new training and testing set splits. The average nMSE is slightly lower and the average wR is slightly higher for the calibrated TGL. Compared to TGL, the calibrated TGL has a lower nMSE on 10 of the 20 splits and a higher wR on 9 of the 20.

nMSE, TGL	nMSE, calib TGL	wR, TGL	wR, calib TGL
Mean: 0.4568	Mean: 0.4555	Mean: 0.7529	Mean: 0.7580
0.4029	0.3930	0.7678	0.7721
0.4313	0.4826	0.7612	0.7408
0.5357	0.5976	0.7063	0.7994
0.5552	0.4359	0.7075	0.7586
0.4311	0.5094	0.7609	0.7083
0.3518	0.5058	0.8126	0.7205
0.5788	0.4158	0.6865	0.7808
0.5045	0.4118	0.7251	0.8223
0.4494	0.4777	0.7492	0.7340
0.4292	0.3965	0.7640	0.8005
0.4899	0.4041	0.7396	0.7737
0.4484	0.4311	0.7352	0.7583
0.3956	0.3892	0.7951	0.7841
0.5348	0.4570	0.7555	0.7529
0.4092	0.4741	0.7790	0.7473
0.3719	0.4186	0.8175	0.7708
0.4303	0.4965	0.7680	0.7347
0.4405	0.5104	0.7495	0.7160
0.4457	0.4899	0.7632	0.7199
0.4989	0.4114	0.7126	0.7644

model parameters but train and test on 20 new splits of the data to obtain results shown in Table 4.5 and Table 4.6. On these new splits, the calibrated models perform equally as well as the TGL and cFSGL.

Table 4.6: Comparison of MMSE prediction from calibrated versus non-calibrated cFSGL on 20 new training and testing set splits. The average nMSE is lower and the average wR higher for the calibrated cFSGL. Compared to cFSGL, the calibrated cFSGL has a lower nMSE on 10 of the 20 splits and a higher wR on 10 of the 20.

nMSE, cFSGL	nMSE, calib cFSGL	wR, cFSGL	wR, calib cFSGL
Mean: 0.4596	Mean: 0.4506	Mean: 0.7528	Mean: 0.7598
0.3958	0.3918	0.7688	0.7741
0.4324	0.4827	0.7579	0.7469
0.5517	0.5778	0.7046	0.7957
0.5722	0.4356	0.7004	0.7563
0.4434	0.4998	0.7620	0.7146
0.3450	0.4998	0.8155	0.7198
0.5921	0.4175	0.6788	0.7830
0.4869	0.4021	0.7334	0.8279
0.4675	0.4769	0.7368	0.7347
0.4193	0.3987	0.7724	0.7998
0.4820	0.4075	0.7417	0.7727
0.4727	0.4306	0.7340	0.7565
0.3976	0.3763	0.7914	0.7880
0.5718	0.4558	0.7427	0.7522
0.4005	0.4713	0.7813	0.7490
0.3517	0.4173	0.8277	0.7718
0.4354	0.4898	0.7611	0.7405
0.4309	0.5033	0.7610	0.7174
0.4442	0.4814	0.7657	0.7233
0.4987	0.4016	0.7180	0.7711

Chapter 5

Discussion

Our purpose in this study is to conduct a thorough comparison of regression techniques, both well-established and novel, on a limited set of high-dimensional data. If our goal is to be able to predict cognitive scores that are even more closely correlated to the truth, other demographic factors such as age (the number one risk factor for AD according to Association) should be added as an input feature. It would also be valuable to consider additional information available on ADNI, like race and APOE $\epsilon 4$ genotype. There is conflicting evidence in the literature as to whether those other demographic features are useful - Zhou et al. (2013) found that the addition of these features greatly improved the correlation coefficient at all time points, while Murphy et al. (2010) claimed that they were not significant in the multivariate regression.

Our first result presented in this thesis is the derivation of the smooth approximation of the calibrated TGL and calibrated cFSGL models. The second is an experimental result, namely finding that TGL performs slightly better than cFSGL on predicting MMSE scores for the ADNI1 cohort. This contradicts the result presented in Zhou et al. (2013), albeit for a different data set. A possible explanation for this is that the ADNI1 we worked with may be more homogenous, so the restrictive penalty in TGL might be more appropriate than the smooth penalty in cFSGL. The

ADNI2 data, meanwhile, consists of a wider range of patients (five levels of diagnosis rather than three), and with a larger number of features, the sparsity inducing regularizations of the cFSGL may play a more important role in selecting features. The ADNI2 data for MMSE also has both more available values than the data from Zhou et al. (2013) for earlier time points, but fewer available values for later time points, which could explain the sharp drop in correlation between predicted and true target values, which was not observed in their results.

The design of ADAS-Cog is such that a higher score reflects a higher measure of cognitive decline, while in MMSE a higher score represents cognitive function that is closer to normal. Across all time points in Part II of our study, the tendency of by the model built from the cFSGL formulation is to overpredict the mental cognition for patients who had a low true MMSE score, while the error for patients who had a high true score was much smaller. This may be due to much fewer numbers of patients with AD than numbers of CN patients included in the data set, which makes training coefficient values for those types of subjects harder.

In the ADNI2 phase, the number of times for which some patients have MMSE or ADAS-Cog 13 scores is far greater than five and extend up to M96. Further work could be done to explore the predictive power of these models when give more tasks. The idea behind multi-task learning is an advantage over single-task methods when there are many tasks but few data points for each. So, including time points M18, M60, and more, even if scores are available only for a small percentage of patients, may reveal the models to be more useful than presented in our results here.

Our final result was initial evidence that the calibrated multi-task models performed worse on our ADNI2 data set than their non-calibrated counterparts, despite their theoretical advantages. However, this may have been due to the condition by which we were selecting the optimal model parameters, which was to minimize the nMSE, which is similar in form to the loss function in the traditional multi-task

models. By reporting the performance measures from the same sets that we were fitting parameters to, the process would certainly favor the model similar to the very expression that we attempted to minimize. Therefore, we trained and tested the models on 20 new random splits of data, and found almost equal performance between non-calibrated and calibrated models. On the same split of data, one method would sometimes greatly outperform the other, suggesting the calibrated formulations have some advantage when the training data meets certain conditions.

Chapter 6

Conclusion

In this study, we first analyzed the performance of the single-task ridge and Lasso methods against the performance of the multi-task temporal group Lasso (TGL) and convex fused sparse group Lasso (cFSGL) methods on two separate data sets. In the first, consisting of features from 1.5T ADNI1 MRI images, the TGL formulation outperformed the three other methods for the prediction of MMSE scores, while both multi-task models performed well for the prediction of ADAS-Cog scores. cFSGL would consistently have a lower root mean squared error for the prediction of scores at later time points, likely due to its three non-smooth regularizations.

In experiments on the second data set, consisting of features from 3T ADNI2 MRI images, the cFSGL formulation performed better by all measures of predictive performance. We then built linear prediction models trained on 75% of the MMSE data and tested the performance on the remaining 25% by regressing the predicted target value on the ground truth. Unsurprisingly, for time points further into the future, the correlation between predicted and true MMSE score decreased. Across all time points, however, the tendency was to overpredict the mental cognition ability for patients whose true mental cognition ability was low, while the error for patients who were close to cognitively normal was much smaller.

Lastly, we formulated, implemented and analyzed the performance of three novel multi-task learning methods for the prediction of AD progression at different time points, utilizing the modified loss function from the calibrated multivariate model. Though average performance over many different splits of data was similar between TGL and calibrated TGL and between cFSGL and calibrated cFSGL, further experimentation showed that on the same split of the data, the difference in performance between calibrated and non-calibrated formulation could be quite large. Therefore, it is likely that by discriminating on the types of input data, namely those that satisfy the conditions in Liu et al. (2014), these novel calibrated multi-task models would outperform the multi-task models in today's literature when it comes to predicting the progression of AD.

Appendix A

Code

Code for Lasso and ridge regression techniques are readily available in R statistical packages. In this thesis we utilized the `glmnet` package (Friedman et al., 2010). For the calibrated multi-task models, the code for finding W under the calibrated TGL and the calibrated cFSGL formulations can be found below.

```
function [ W, info ] = calibTGL_SPG( X, y, lambda1, lambda2, ...
    lambda3, opts )
%
% Multi-Task Feature Learning with Calibration - Smoothing Proximal
% Gradient (SPG)
%
% diagonalized
%
% OBJECTIVE
%     min_W { sum_i^m ||Xi wi - yi||_mu + lambda1 ||W||_{1,2} +
%                 lambda2/2 ||W||_F^2  + lamda3 ||WH||_F^2}
%         ||Xi wi - yi||_mu = max <vi, Xi wi - yi> s.t. ||vi||<=1.
%
%
% INPUT
%  X - cell array of {n_i by d matrices} by m
%  y - cell array of {n_i by 1 vectors}  by m
%  lambda1 - regularization parameter of the l2,1 norm penalty
%  lambda2 - regularization parameter of the Fro norm penalty
%  lambda3 - regularization parameter of the temporal smoothness penalty
%
% OUTPUT
%  W - task weights: d by t.
%  funcVal - the funcion value.
```

```
%
% MALSAR authors: Jiayu Zhou, Pinghua Gong
% Modified by: Yuan Cao, Rebecca Zhang

%% Initialization
if(nargin<6), opts = []; end

% 0 nothing. 1 iteration num. 2 iteration details.
opts = setOptsDefault( opts, 'verbose', 1);
opts = setOptsDefault( opts, 'maxIter', 10000);
opts = setOptsDefault( opts, 'tol',     1e-7);
opts = setOptsDefault( opts, 'epsilon', 1e-1);
opts = setOptsDefault( opts, 'stopflag', 1);
verbose = opts.verbose;

info.algName = 'Smth SpaRSA';

if verbose > 0
    fprintf('%s Config [MaxIter %u][Tol %.4g][Epsilon %.4g]\n', ...
        info.algName, opts.maxIter, opts.tol, opts.epsilon);
end

m = length(X); % task number
d = size(X{1}, 2);

H=zeros(m,m-1);
H(1:(m+1):end)=1;
H(2:(m+1):end)=-1;

% diagonalize X and vectorized y.
[Xdiag, ~, Th_vecIdx, yvect] = diagonalize(X, y);

sigma = 1e-3;    % line search constant.
eta_init = 1;    % initial value of eta (1 = beta^0).
eta_max  = 1e+14;
eta_min  = 1e-14;
beta = 1.15;     % eta incremental
mu =  opts.epsilon/m;
muInc = 1;     % decrease of mu.

if isfield(opts, 'initW')
    W0 = opts.initW;
    if verbose > 0, fprintf('%s: use given initial point.\n', ...
       info.algName), end
else
    W0 = randn(d, m); % starting from a random point.
end

info.fvP = zeros(opts.maxIter, 1);
funcVal  = zeros(opts.maxIter, 1);
timeVal  = zeros(opts.maxIter, 1);

lsCnt = 0; % count for line search.
```

```
%% Computation
if verbose == 1; fprintf('Iteration:     '); end

Wk = W0;

for iter = 1: opts.maxIter
    iterTic = tic;

    [ fWk_mu , gWk_mu] = smoothObjective(Wk, mu);

    eta = eta_init;
    if iter >1  % BB-rule.
        xx = Wk - Wk_old; yy = gWk_mu - gWk_mu_old;
        eta = sum(sum(xx .* yy))/sum(sum(xx .* xx));
        eta = min(eta_max, max(eta_min, eta));
    end
    for lsIter = 1:100
        % smooth projection
        Wk_new = proj(Wk - gWk_mu / eta, lambda1/eta);
        [fWk_new_mu, XWy, thNrms ] = smoothObjectiveFv(Wk_new, mu);

        % line search
        if (fWk_mu - fWk_new_mu) >= sigma * eta /2 * sum(sum((Wk_new ...
            - Wk).^2))
            break;
        end

        eta = eta * beta;
    end
    lsCnt = lsCnt + lsIter;

    Wk_old     = Wk;
    gWk_mu_old = gWk_mu;
    Wk = Wk_new;

    funcVal(iter)  = fWk_new_mu; % the smoothed objective function
    info.fvP(iter) = primalObjective(XWy, thNrms); % the primal ...
       function.

    if iter > 1, timeVal(iter) = timeVal(iter-1) + toc(iterTic);
    else timeVal(iter) = toc(iterTic); end

    if verbose == 1; fprintf('\b\b\b\b\b%5i',iter); end
    if verbose ==2
        fprintf('%s: [Iteration %u][fvP %.4g][fvS %.4g][LS %u]\n', ...
            info.algName, iter, info.fvP(iter), funcVal(iter), lsCnt);
    end

    % test stop condition.
    if iter>=2
        switch opts.stopflag
            case 1
                if (abs( funcVal(iter) - funcVal(iter-1) ) <=...
                        opts.tol* abs(funcVal(iter-1)))
```

```
                    break;
                end
            case 2
                if ~isfield(opts, 'obj')
                    error('opts.obj must be set');
                end
                if abs(info.fvP(iter) - opts.obj) <= opts.tol*opts.obj
                    break;
                end
            case 3
                if ~isfield(opts, 'W')
                    error('opts.W must be set');
                end
                if norm(Wk - opts.W,'fro') <= ...
                   opts.tol*norm(opts.W,'fro')
                    break;
                end
        end
    end

    mu = mu * muInc;
end
if verbose == 1; fprintf('\n'); end

%% Output.
W    = Wk;
info.funcVal = funcVal (1:iter);
info.fvP     = info.fvP(1:iter);
info.timeVal = timeVal (1:iter);


%% Nested Functions
    function fv = primalObjective(XWy, thNrms)
        % primal objective
        %  P(W)  sum_i^m ||Xi wi - yi|| + lambda1 ||W||_{1,2} + ...
           lambda2/2 ||W||_F^2
        %                     + lambda3 ||WH||_F^2
        fv =  segL2 (XWy, Th_vecIdx) + thNrms;
    end

    function [f, g] = smoothObjective(W, mu)
        % f: funcVal of the ENTIRE smoothing function (including l21).
        % g: gradient of the smooth part of the smoothing function.

        XWy = Xdiag * W(:) - yvect;
        vv = segL2Proj (XWy/mu, Th_vecIdx);
        f = lambda2 / 2 * sum(sum(W.^2))  + lambda1 * L21norm(W) ...
            + lambda3 * sum(sum((W*H).^2))+ vv' * XWy - mu/2 * ...
               sum(vv.^2);

        g = lambda2 * W + 2*lambda3*(W*(H*H')) + reshape( Xdiag' * ...
           vv, size(W));
    end
```

```
    function [f, XWy, thNrms] = smoothObjectiveFv(W, mu)
        % f: funcVal of the ENTIRE smoothing function (including l21).
        thNrms = lambda2 / 2 * sum(sum(W.^2))  + lambda1 * ...
            L21norm(W) ...
             + lambda3 * sum(sum((W*H).^2));

        % NOTE:there is a 1e-15 difference from matlab results on ...
            gradient.
        XWy = Xdiag * W(:) - yvect;
        vv = segL2Proj (XWy/mu, Th_vecIdx);
        f = thNrms + vv' * XWy - mu/2 * sum(vv.^2);
    end

end

function [Xnrm] = L21norm(X)
% ||X||_{1,2} = sum_i ||X^i||_2
Xnrm = sum(sqrt(sum(X.^2, 2)));
end

function [X] = proj(D, lambda )
% l2.1 norm projection.
X = repmat(max(0, 1 - lambda./sqrt(sum(D.^2,2))),1,size(D,2)).*D;
end


function [ W, info ] = calibcFSGL_SPG( X, y, lambda1, lambda2, ...
    lambda3, opts )
%
% Multi-Task Feature Learning with Calibration - Smoothing Proximal
% Gradient (SPG)
%
% diagonalized
%
% OBJECTIVE
%    min_W { sum_i^m ||Xi wi - yi||_mu + lambda1 ||W||_{1,2} +
%              lambda2 ||W||_1  + lambda3 ||WH||_1}
%        ||Xi wi - yi||_mu = max <vi, Xi wi - yi> s.t. ||vi||<=1.
%
%
% INPUT
%  X - cell array of {n_i by d matrices} by m
%  y - cell array of {n_i by 1 vectors}  by m
%  lambda1 - regularization parameter of the l2,1 norm penalty
%  lambda2 - regularization parameter of the l1 Lasso norm penalty
%  lambda3 - regularization parameter of the fused Lasso penalty
%
% OUTPUT
%  W - task weights: d by t.
%  funcVal - the function value.
%
% MALSAR authors: Jiayu Zhou, Pinghua Gong
% Modified by: Yuan Cao, Rebecca Zhang
```

```
%% Initialization
if(nargin<6), opts = []; end

% 0 nothing. 1 iteration num. 2 iteration details.
opts = setOptsDefault( opts, 'verbose', 1);
opts = setOptsDefault( opts, 'maxIter', 10000);
opts = setOptsDefault( opts, 'tol',     1e-7);
opts = setOptsDefault( opts, 'epsilon', 1e-1);
opts = setOptsDefault( opts, 'stopflag', 1);
verbose = opts.verbose;

info.algName = 'Smth SpaRSA';

if verbose > 0
    fprintf('%s Config [MaxIter %u][Tol %.4g][Epsilon %.4g]\n', ...
        info.algName, opts.maxIter, opts.tol, opts.epsilon);
end

m = length(X); % task number
d = size(X{1}, 2);

H1=zeros(m,m-1);
H1(1:(m+1):end)=1;
H1(2:(m+1):end)=-1;
H = [lambda2*eye(m) lambda3*H1];

% diagonalize X and vectorized y.
[Xdiag, ~, Th_vecIdx, yvect] = diagonalize(X, y);

sigma = 1e-3;    % line search constant.
eta_init = 1;    % initial value of eta (1 = beta^0).
eta_max  = 1e+14;
eta_min  = 1e-14;
beta = 1.15;     % eta incremental
mu =  opts.epsilon/m;
muInc = 1;     % decrease of mu.

if isfield(opts, 'initW')
    W0 = opts.initW;
    if verbose > 0, fprintf('%s: use given initial point.\n', ...
       info.algName), end
else
    W0 = randn(d, m); % starting from a random point.
end

info.fvP = zeros(opts.maxIter, 1);
funcVal  = zeros(opts.maxIter, 1);
timeVal  = zeros(opts.maxIter, 1);

lsCnt = 0; % count for line search.

%% Computation
if verbose == 1; fprintf('Iteration:     '); end
```

```
Wk = W0;

for iter = 1: opts.maxIter
    iterTic = tic;

    [ fWk_mu , gWk_mu] = smoothObjective(Wk, mu);

    eta = eta_init;
    if iter >1  % BB-rule.
        xx = Wk - Wk_old; yy = gWk_mu - gWk_mu_old;
        eta = sum(sum(xx .* yy))/sum(sum(xx .* xx));
        eta = min(eta_max, max(eta_min, eta));
    end
    for lsIter = 1:100
        % smooth projection
        Wk_new = proj(Wk - gWk_mu / eta, lambda1/eta);
        [fWk_new_mu, XWy, thNrms ] = smoothObjectiveFv(Wk_new, mu);

        % line search
        if (fWk_mu - fWk_new_mu) >= sigma * eta /2 * sum(sum((Wk_new ...
            - Wk).^2))
            break;
        end

        eta = eta * beta;
    end
    lsCnt = lsCnt + lsIter;

    Wk_old     = Wk;
    gWk_mu_old = gWk_mu;
    Wk = Wk_new;

    funcVal(iter)  = fWk_new_mu; % the smoothed objective function
    info.fvP(iter) = primalObjective(XWy, thNrms, Wk); % the primal ...
       function.

    if iter > 1, timeVal(iter) = timeVal(iter-1) + toc(iterTic);
    else timeVal(iter) = toc(iterTic); end

    if verbose == 1; fprintf('\b\b\b\b\b%5i',iter); end
    if verbose ==2
        fprintf('%s: [Iteration %u][fvP %.4g][fvS %.4g][LS %u]\n', ...
            info.algName, iter, info.fvP(iter), funcVal(iter), lsCnt);
    end

    % test stop condition.
    if iter>=2
        switch opts.stopflag
            case 1
                if (abs( funcVal(iter) - funcVal(iter-1) ) <=...
                        opts.tol* abs(funcVal(iter-1)))
                    break;
                end
            case 2
```

```
                if ~isfield(opts, 'obj')
                    error('opts.obj must be set');
                end
                if abs(info.fvP(iter) - opts.obj) <= opts.tol*opts.obj
                    break;
                end
            case 3
                if ~isfield(opts, 'W')
                    error('opts.W must be set');
                end
                if norm(Wk - opts.W,'fro') <= ...
                   opts.tol*norm(opts.W,'fro')
                    break;
                end
        end
    end

    mu = mu * muInc;
end
if verbose == 1; fprintf('\n'); end

%% Output.
W    = Wk;
info.funcVal = funcVal (1:iter);
info.fvP     = info.fvP(1:iter);
info.timeVal = timeVal (1:iter);


%% Nested Functions
    function fv = primalObjective(XWy, thNrms, W)
        % primal objective
        %  P(W)  sum_i^m ||Xi wi - yi|| + lambda1 ||W||_{1,2} + ...
           lambda2/2 ||W||_F^2
        %                     + lambda3 ||WH||_F^2
        fv =  segL2 (XWy, Th_vecIdx) + thNrms + sum(sum(abs(W*H)));
    end

    function [f, g] = smoothObjective(W, mu)
        % f: funcVal of the ENTIRE smoothing function (including l21).
        % g: gradient of the smooth part of the smoothing function.

        XWy = Xdiag * W(:) - yvect;
        vv = segL2Proj (XWy/mu, Th_vecIdx);
        % optimal A for fused Lasso penalty
        A = (W*H)/mu;
        A(A > 1) = 1;
        A(A < -1) = -1;
        f = lambda1 * L21norm(W) + vv' * XWy - mu/2 * sum(vv.^2) ...
            + sum(sum((W*H).*A)) - mu/2 * sum(sum(A.^2));

        g = reshape( Xdiag' * vv, size(W)) + A*H';
    end

    function [f, XWy, thNrms] = smoothObjectiveFv(W, mu)
```

```
        % f: funcVal of the ENTIRE smoothing function (including l21).
        thNrms = lambda1 * L21norm(W);

        % NOTE:there is a 1e-15 difference from matlab results on ...
           gradient.
        XWy = Xdiag * W(:) - yvect;
        vv = segL2Proj (XWy/mu, Th_vecIdx);
        A = (W*H)/mu;
        A(A > 1) = 1;
        A(A < -1) = -1;

        f = thNrms + vv' * XWy - mu/2 * sum(vv.^2) + ...
            sum(sum((W*H).*A)) - mu/2 * sum(sum(A.^2)) ;
    end

end

function [Xnrm] = L21norm(X)
% ||X||_{1,2} = sum_i ||X^i||_2
Xnrm = sum(sqrt(sum(X.^2, 2)));
end

function [X] = proj(D, lambda )
% l2.1 norm projection.
X = repmat(max(0, 1 - lambda./sqrt(sum(D.^2,2))),1,size(D,2)).*D;
end
```

Subfunctions `segL2` and `segL2Proj` are written in C and can be found in the MALSAR package.

Bibliography

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Mach Learn*, 73:243–272.

Association, A. (2015). 2015 alzheimer's disease facts and figures. Technical report, Alzheimer's Association.

Caroli, A. and Frisoni, G. (2010). The dynamics of alzheimer's disease biomarkers in the alzheimer's disease neuroimaging initiative cohort. *Neurobiol Aging*, 31(8):1263–1274.

Chetelat, G. and Baron, J. (2002). Early diagnosis of alzheimer's disease: contribution of structural neuroimaging. *NeuroImage*, 18:525–541.

Duchesne, S., Caroli, A., Geroldi, C., Collins, D. L., and Frisoni, G. (2009). Relating one-year cognitive change in mild cognitive impairment to baseline mri features. *NeuroImage*, 47:1363–1370.

Evgeniou, T., Micchelli, C., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frisoni, G., Testa, C., Zorzan, A., Sabattoli, F., Beltramello, A., Soininen, H., and Laakso, M. (2002). Detection of grey matter loss in mild alzheimer's disease with voxel based morphometry. *Journal of Neurology, Neurosurgery and Psychiatry.*

Ito, K., Corrigan, B., Zhao, Q., French, J., Miller, R., Soares, H., Katz, E., Nicholas, T., Billing, B., Anziano, R., and Fullerton, T. (2011). Disease progression model for cognitive deterioration from alzheimer's disease neuroimaging initiative database. *Alzheimers Dement*, 7(2):151–160.

Liu, H., Wang, L., and Zhao, T. (2014). Multivariate regression with calibration. In *Advances in Neural Information Processing Systems 27.*

Murphy, E., Holland, D., Donohue, M., McEvoy, L., Hagler, D., Dale, A., and Brewer, J. (2010). Six-month atrophy in mtl structures is associated with subsequent memory decline in elderly controls. *NeuroImage*, 53(4):1310–1317.

Petrella, J., Coleman, R., and Doraiswamy, P. (2003). Neuroimaging and early diagnosis of alzheimer disease: a look to the future. *Radiology*.

Stonnington, C., Chu, C., Klöppel, S., Jack, C., Ashburner, J., and Frackowiak, R. (2010). Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *NeuroImage*, 51(4):1405–1413.

Thompson, P., Hiyashi, K., de Zubicaray, G., Janke, A., Rose, S., Semple, J., Hong, M., Herman, D., Gravano, D., Doddrell, D., and Toga, A. (2004). Mapping hippocampal and ventricular change in alzheimer disease. *NeuroImage*, 22:1754–1766.

Thrun, S. (1996). *Is learning the n-th thing any easier than learning the first?* The MIT Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society.*, 58(1):267–288.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society.*, 68(1):49–67.

Zhou, J., Chen, J., and Ye, J. Malsar: Multi-task learning via structural regularization.

Zhou, J., Chen, J., and Ye, J. (2011). Clustered multi-task learning via alternating structure optimization. *Advances in Neural Information Processing Systems*, pages 702–710.

Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2013). Modeling disease progression via multi-task learning. *NeuroImage*, 78:233–248.